# Calculation of spectral similarity independent of measurement equipment

Ryo Murakami[1,a) Hiroshi Shinotsuka[2] Kenji Nagata[2] Hideki Yoshikawa[2]
Hayaru Shouno[1,b)]

**Abstract:** We often compared measured photoelectron spectra with other spectra for material development and quality control in the industry. In particular, X-ray photoelectron spectroscopy is used to detect surface contamination and chemical state changes. However, spectral data has perturbation of measurement devices, e.g., the difference in peak width due to the resolution of the device, and the difference in peak position due to the charging phenomenon in the spectral data. It is difficult to simply measure the distance between the measured spectra. Therefore, it is necessary to develop a method for calculating the similarity between spectra that is independent of the device. To establish a comparing procedure, we introduced a clustering method for spectral data to decouple the measurement perturbation. We designed the clustering method for detecting contamination components and sample heterogeneity. This study proposed an analytical model that separates the photoelectron peaks from the perturbation caused by the measurement device. We applied the method to calculate the similarity between the spectra. As a result, we show the proposed method could detect spectral data included with other components in the analysis of real X-ray photo-electron spectroscopy spectral data of $TiO_2$.

**Keywords:** Spectral analysis, X-ray photo-electron spectroscopy, Removal of perturbations, Peak separation

## 1. Introduction

Photoelectron spectroscopy is a method of measuring the electronic state of a sample by emitting an electromagnetic wave of certain energy and measuring the energy of the electrons emitted by the photoelectric effect. It is used for material development and quality control in the industry. In particular, X-ray photoelectron spectroscopy (XPS) measures the kinetic energy distribution of photoelectrons emitted by X-ray irradiation. XPS has attracted attention because it can perform surface analysis. XPS can measure many materials, including metallic and polymeric materials. In recent years, analysis methods have been actively proposed with the demand for automated analysis [1, 2]. The methods of Nagata et al. and Shinozuka et al. can automatically estimate the number of peaks and the parameters of the peaks [1, 2]. In a related study, we proposed the analysis method that can automatically estimate compound ratios using reference spectra obtained from the literature [3]. This method has succeeded in significantly reducing the variability of the analytical results. Meanwhile, it is not established how to measure the distances between the spectra and how the spectra can be clustered.

In material development, similarity of intrinsic spectral structures is essential for discussing material properties and quality control in the industry. Furthermore, clustering of XPS spectra based on their similarity can be applied to the detection of surface state changes. In spectral analysis, we often compare measurement spectral data with other spectral data for material development and quality control. However, spectral data has perturbation of measurement devices, e.g., the difference in peak width due to the resolution of the device, and the difference in peak position due to the charging phenomenon in the spectral data. Therefore, it is challenging to compare the spectral measurement data with the data from the literature and the spectral data measured by equipment of other institutions. The perturbations should be modeled from the measurement equipment, and methods developed for comparing spectra independent of the perturbations. In particular, the task of this study is the clustering of XPS spectral data independent of the measurement device.

We designed the clustering to detect contamination components and sample heterogeneity for material development and industrial quality control. This study introduced an analytical model that separates the photoelectron peaks and the perturbation caused by the measurement device. We proposed a method to calculate the similarity between spectra. This study shows the proposed technique's effectiveness from its application to artificial data and real XPS spectral data.

## 2. Method

In this study, we aimed to develop a clustering method for XPS spectral data independent of the measurement equipment. It is important to remove device-induced fluctuations from the

[1] The University of Electro-Communications, Graduate School of Informatics and Engineerings, Chofu 182-8585, Tokyo, Japan
[2] Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba 305-0044, Japan
[a)] m2240010@edu.cc.uec.ac.jp
[b)] shouno@uec.ac.jp

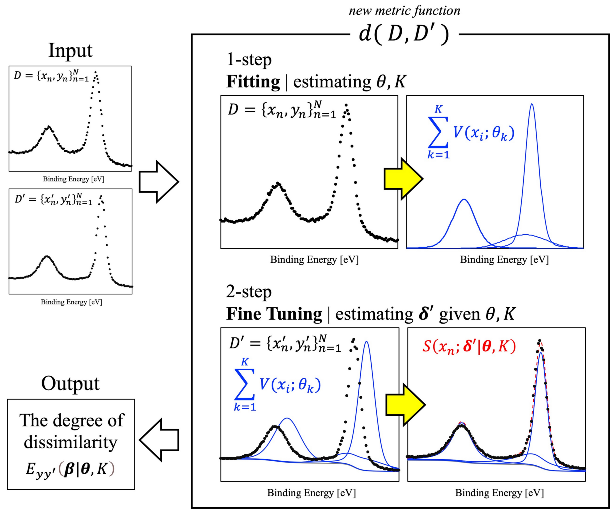
new metric function
$d(D, D')$
Input
$D = \{x_n, y_n\}_{n=1}^N$
Binding Energy [eV]
$D' = \{x'_n, y'_n\}_{n=1}^N$
Binding Energy [eV]
1-step
Fitting | estimating $\theta, K$
$D = \{x_n, y_n\}_{n=1}^N$
$\sum_{k=1}^K V(x_i; \theta_k)$
Binding Energy [eV]
Binding Energy [eV]
2-step
Fine Tuning | estimating $\boldsymbol{\delta}'$ given $\theta, K$
$D' = \{x'_n, y'_n\}_{n=1}^N$
$\sum_{k=1}^K V(x_i; \theta_k)$
$S(x_n; \boldsymbol{\delta}'|\boldsymbol{\theta}, K)$
Binding Energy [eV]
Binding Energy [eV]
Output
The degree of dissimilarity
$E_{yy'}(\boldsymbol{\beta}|\boldsymbol{\theta}, K)$

**Fig. 1** Conceptual diagram of the proposed method that compute degree of dissimilarity in spectrum-to-spectrum excluding change in spectral shape depending on measurement device.

spectrum. In this study, the key concept is to compute a degree of dissimilarity in spectrum-to-spectrum, excluding change in spectral shape depending on the measurement device. The proposed method calculated the degree of dissimilarity in spectrum-to-spectrum by two steps:

- Step 1:
  We calculate the fitting peak of the spectral data $D = \{(x_i, y_i)\}_{i=1}^N$ throughout a pseudo-Voigt function. In this step, we also estimate the number of peaks $K$ and the peak parameters $\theta = \{\boldsymbol{h}, \boldsymbol{p}, \boldsymbol{w}, \boldsymbol{r}\}$.
- Step 2:
  We carry out fine-tuning on the estimated parameters $\theta$ to apply them to other spectral data $D' = \{(x'_i, y'_i)\}_{i=1}^{N'}$.

In step 1, we applied a peak separation method proposed by Shinotsuka et al [2]. Their method, which is based on the Bayesian information criterion (BIC) [4], allowed us to estimate the peak parameters and the number of peaks automatically. Step 2 means adjusting the change in spectral shape depending on the measurement device. After these steps, the proposed method calculated the error (sum of squared residuals) between spectral data $y'$ and the fitted function based on peak parameters after fine-tuning. The calculated error shows how well the fitted model (the number of peaks $K$ and the peak parameters $\theta$) based on a spectral data $D$ can explain another spectral data $D'$ by adjusting the perturbation due to the measurement device. That is, the calculated error corresponds to the degree of dissimilarity between $D$ and $D'$. Hereafter, we denote it as the function $d(D, D')$ in this study. $d(D, D')$ excludes the change in spectral shape depending on the measurement device. This means that $d(D, D')$ is a dissimilarity of the spectra based on the structure of the photoelectron peak derived from the measured sample.

The proposed method applied the dissimilarity function $d(D, D')$ to all pairs of $M$ spectral data. As a result, we can obtain that we can obtain the dissimilarity matrix $G$. Therefore, $G$ consists of the following elements:

$$G_{ij} = d(D_i, D_j), \tag{1}$$

where $G \in R^{M \times M}$, $G_{ij} \geq 0$. Then we apply the principal component analysis (PCA) to the dissimilarity matrix $G$ for visualization. We used the computed principal component (refer to as PC) to detect the contamination spectral data.

### 2.1 Formulation of Fine Tuning

This section describes the fine-tuning method to adjust variations in spectral shape depending on measurement device. Here, we assume that $\{x, y\}$ and $\{x', y'\}$ are the spectral data of the same sample measured in different measurement devices. Therefore, $\{x, y\}$ and $\{x', y'\}$ are described by the same peak structure $\theta$ and the number of peaks $K$. However, the peak structure is perturbed due to differences in the measurement devices. Here, we define the peak parameters $\boldsymbol{\theta} = \{h_k, p_k, w_k\}_{k=1}^K$ as follows:

- $h_k$: peak intensity
- $p_k$: peak position
- $w_k$: peak width

In this study, the perturbations are described as parameters $\delta$. In Fine Tuning, we optimize $\delta$ under the given $\{x', y'\}$ and $\theta$. First, in Step 1, we estimated the parameters $\{K, \theta\}$ from $\{x, y\}$. Given the parameters $\{K, \theta\}$, the proposed method applied fine-tuning to other spectral data $y' = \{y'_n\}_{n=1}^{N'}$ as follows:

$$\begin{aligned} y'_n &\approx f_{K,\theta}(x'_n; \boldsymbol{\delta}', a, b), & (2) \\ &\approx S_{K,\theta}(x'_n; \boldsymbol{\delta}') + B(x'_n; a, b), & (3) \end{aligned}$$

where $f_{K,\theta}(x'_n; \boldsymbol{\delta}', a, b)$ is the fitting function. $S_{K,\theta}(x'_n; \boldsymbol{\delta}')$ and $B(x'_n; a, b)$ mean the signal spectrum and the background, respectively. In the calculation of the background $B(x_n; a, b)$, we apply Shirley method [5, 6]. The parameter $x'_n$ means the binding energy. Here, we define the adjusting components $\boldsymbol{\delta}' = \{\eta, \mu, \omega, \{r_k\}_{k=1}^K\}$ as follows:

- $\eta$: intensity adjustment
- $\mu$: position adjustment
- $\omega$: width adjustment
- $\{r_k\}_{k=1}^K$: Lorentz-Gauss ratio (LG ratio) of each peak
- $\{a, b\}$: End-point intensity of the background

We assume that a sum of the pseudo-Voigt functions composes the signal spectrum $S_{K,\theta}(x'_n; \delta')$ as:

$$S_{K,\theta}(x'_n; \boldsymbol{\delta}') = \sum_{k=1}^{K} \eta h_k V(x_n; p_k + \mu, \omega w_k, r_k). \tag{4}$$

The pseudo-Voigt function $V(x_n; p, w, r)$ has three parameters $p$, $w$, and $r$ [7], where each parameter means the center peak position, width and LG-ratio, respectively. In fine-tuning, the estimating parameters is $\boldsymbol{\beta} = \{\boldsymbol{\delta}', a, b\}$, and determine $\boldsymbol{\beta}$ so that the error function $E_{yy'}(\boldsymbol{\beta}|\boldsymbol{\theta}, K)$ was minimized. The error function $E_{yy'}(\boldsymbol{\beta}|\boldsymbol{\theta}, K)$ is defined by the following equation:

$$E_{yy'}(\boldsymbol{\beta}|\boldsymbol{\theta}, K) = \frac{1}{N'} \sum_{n=1}^{N'} \{y'_n - f_{K,\theta}(x'_n; \boldsymbol{\delta}', a, b)\}^2, \tag{5}$$

where $V(x_n; p, w, r)$ is the pseudo Voigt function [7]. Spectral data $y'$ were represented by adjusting a function based on the parameter $\theta$ according to the model of device-derived perturbations. If the error function $E_{yy'}(\boldsymbol{\beta}|\boldsymbol{\theta}, K)$ is high in this analysis model, it means that the intrinsic structure is not similar.

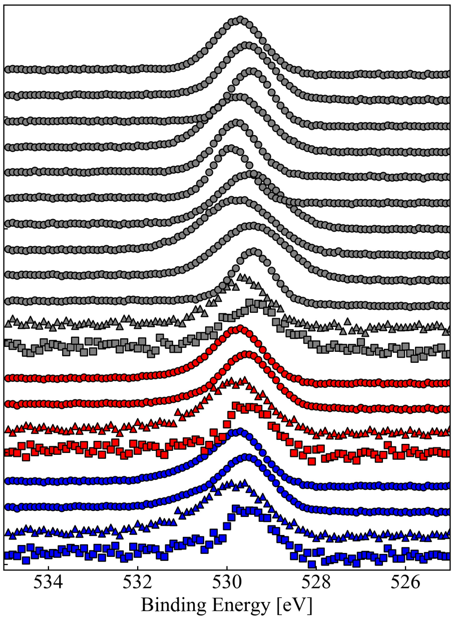

**Fig. 2** 14 artificial spectral data consisting of 10 normal spectral data and 4 spectral data with pseudo surface contamination. Black data are normal, red/blue data are abnormal data with 10/20% contamination, respectively.

The proposed method optimized parameters using the Levenberg-Marquardt method [8] that is one of the gradient methods.

## 3. Results

In order to evaluate the performance of our proposed method, we apply it to both the artificial and real spectral data sets.

### 3.1 Applying to an Artificial Spectral Data

We describe the generation of artificial spectral data. Artificial spectral data were generated on the basis of the pseudo-Voigt function $V(x_n; p_k + \mu, \omega w_k, r_k)$. This study generated 14 artificial spectral data consisting of 10 normal spectral data and 4 spectral data with pseudo surface contamination. The normal spectral data was composed of a single peak with peak position $p_1 = 529.6$ [eV]. In contrast, the abnormal spectral data was composed of summed up the normal spectral data and a single peak with peak position $p_2 = 530.8$ [eV] as pseudo surface contamination.

We added the contamination peak, which has a 10% or 20% area of the main peak area to the normal spectral. Here, we determined that the peak width $w_k$ was 1.0. The peak shift $\mu$, width $\omega$ and LG ratio $r$ by uniform random numbers in the range of $[-0.3, 0.3]$, $[0.5, 1.0]$ and $[0.0, 0.4]$, respectively. This represents a change in the spectral shape depending on the measurement device. Figure 2 shows the artificial spectral data. As shown in Figure 2, the black data is normal and the red / blue data are abnormal data with 10/20% contamination, respectively.

Here, we show results applying three methods to the artificial spectral data. There were two simple methods and the proposed method. Figure 3 shows the scatter plot of PC1 and PC2 when applying PCA to (a) a matrix of raw spectral data, (b) a matrix of the correlation coefficient of the raw spectra $C$ and (c) the dissimilarity matrix proposed by us. The matrix of correlation $C$ consists of the elements $C_{ij} = c(\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(j)})$, where $c(\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(j)})$ is the correlation coefficient of $\boldsymbol{y}^{(i)}$ and $\boldsymbol{y}^{(i)}$. The correlation coefficients are simple metric values that calculate the similarity of the vectors.

In Figure 2, the gray data points are normal spectral data. The red / blue data points are abnormal data with 10/20% contamination. As shown in Figure 3 (a) and (b), the simple method did not allow us to separate normal spectral data (gray data point) and spectral data with pseudo contamination (red / blue data points). This is because we do not consider the change in a spectral shape-derived measurement device. In contrast, as shown in Figure 3 (c), the proposed method can separate normal spectral data and abnormal spectral data with pseudo contamination clearly. Therefore, the proposed method is useful for detecting abnormal spectral data with contamination from multiple XPS spectral data. This investigation shows that the proposed method classifies the XPS spectral data on the basis of the intrinsic peak structure independent of the measurement device.

### 3.2 Applying to the Real Spectral Data

This section describe the real XPS spectral data applying the proposed method. We obtained 15 Ti2p XPS spectral data that are supposed to be pure $TiO_2$ from the literature [9–21]. Figure 4 shows the real XPS spectral data of $TiO_2$ from various literature. As shown in Figure 4, it is confirmed that the variation in peak position and peak shape depending on differences in energy resolution and differences in calibration of the energy axis despite spectral data from the same compound. The purpose of this section is to classify the measured $TiO_2$ spectral data independent of the perturbation component from the measurement device.

We show results applying the proposed method to 15 Ti2p XPS spectral data of $TiO_2$ from literature. Here, PC1 and PC2 can explained about 88% of the dissimilarity matrix $G$ from the explained variance ratio, and thus we consider PC1 and PC2 in this study. Figure 5 shows a scatter plot of PC1 and PC2. In Figure 5, each data point corresponds to each XPS spectral data of $TiO_2$. As shown in Figure 5, it is divided 15 spectral data into two groups: the minority group and the majority group.

## 4. Conclusion

Our purpose was to develop a clustering method for spectral data independent of the measurement device. In this study, we introduced an analytical model that separates the photoelectron peaks and the perturbation due to measurement device, and proposed a method to calculate the similarity between the spectra. The proposed method allowed to classify spectral data based on similarity of intrinsic spectral structures, which is independent of the perturbation component from the measurement device. In this paper, we show two investigations: applying the proposed method to artificial spectral data and the real spectral data. In the artificial spectral data analysis, we shows that the proposed method detected spectral data having pseudo contamination. This clustering method can be expected to be used to detect surface contamination and sample heterogeneity for material development and quality control in industry.

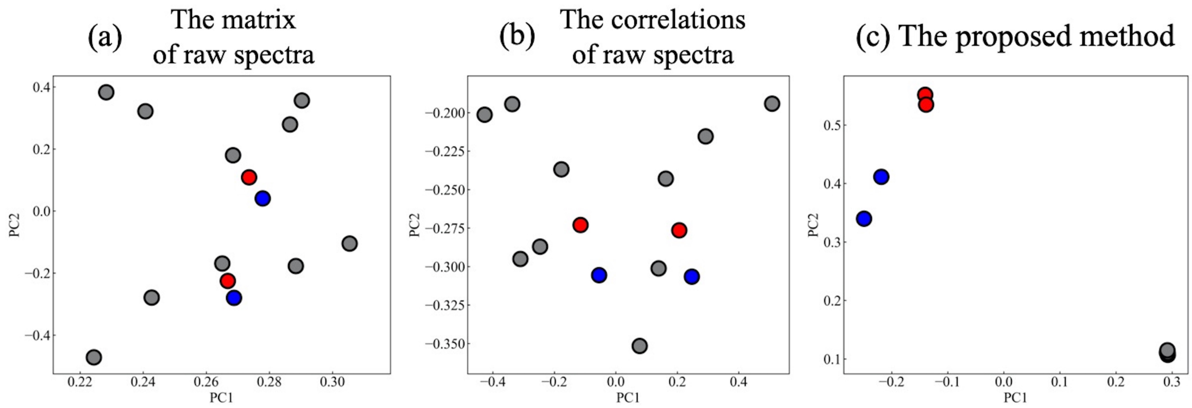

**Fig. 3** Scatter plot of PC1 and PC2 when applying PCA to (a) a matrix of raw spectral data, (b) a matrix of the correlation coefficient of raw spectra $C$ and (c) the dissimilarity matrix calculated by the proposed method. Gray data pints are normal spectral data and red/blue data points are abnormal data with 10/20% contamination respectively.

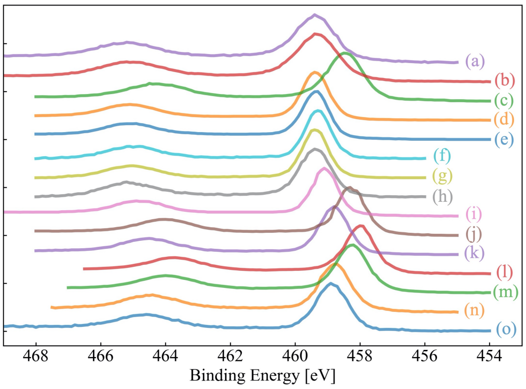

**Fig. 4** Ti 2p XPS spectral data of $TiO_2$ from various literatures. Spectral data (a) to (o) from 15 literature sources [9–21].

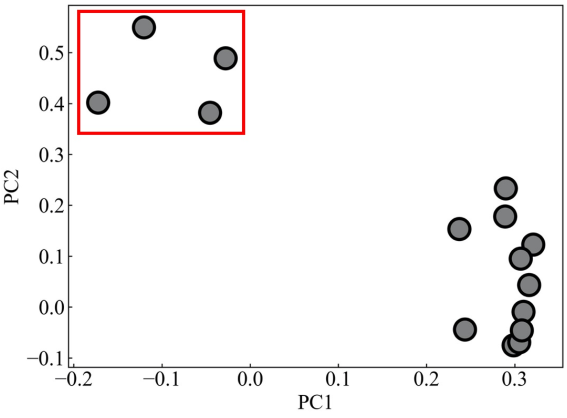

**Fig. 5** Scatter plot of PC1 and PC2 when applying PCA to the dissimilarity matrix calculated by the proposed method in real XPS spectral data of $TiO_2$.

## References

[1] K. Nagata et al., Bayesian spectral deconvolution with the exchange Monte Carlo method. Neural Netw. 28 (2012) 82–89.

[2] H. Shinotsuka, et al., Automated information compression of XPS spectrum using information criteria, J. Electron Spectrosc. Relat. Phenom. 239 (2020) 146903.

[3] R. Murakami, H. Tanaka, H. Shinotsuka et al. Development of multiple core-level XPS spectra decomposition method based on the Bayesian information criterion. J. Electron Spectrosc. Relat. Phenom. 245 (2020) 147003.

[4] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.

[5] D.A. Shirley, High-resolution X-ray photoemission spectrum of the valence bands of gold, Phys. Rev. B 5 (1972) 4709.

[6] A. Proctor, P.M.A. Sherwood, Data analysis techniques in X-ray photoelectron spectroscopy, Anal. Chem. 54 (1982) 13–19.

[7] Peter M.A. Sherwood, Rapid evaluation of the Voigt function and its use for interpreting X-ray photoelectron spectroscopic data, Surface and Interface Analysis 51, 2 (2019) 254–274.

[8] D. W. Marquardt, J. Soc. Indust. Appl. Math. 11, 431 (1963).

[9] D. Ulrike, T. E. Madey, TiO2 by XPS, Surf. Sci. Spectra 4 (1996) 227.

[10] K.D. Schierbaum, S. Fischer, M. C. Torquemada, J.L.de Segovia, E.Román, J.A.Martín-Gago, The interaction of Pt with TiO2(110) surfaces: a comparative XPS, UPS, ISS, and ESD study, Surf. Sci. 345 (1996) 261–273.

[11] M.Z. Atashbar, H. T. Sun, B.Gong, W. Wlodarski, R. Lamb, XPS study of Nb-doped oxygen sensing tio2 thin films prepared by sol-gel method, Thin Solid Films 326 (1998) 238–244.

[12] Bedri Erdem, Robert A. Hunsicker, Gary W. Simmons, E. David Sudol, Victoria L. Dimonie, and Mohamed S. El-Aasser, XPS and FTIR Surface Characterization of TiO2 Particles Used in Polymer Encapsulation, Langmuir 17, 9 (2001) 2664–2669.

[13] Guangming Liu, W. Jaegermann, Jianjun He, Villy Sundström, Licheng Sun, XPS and UPS Characterization of the TiO2/ZnPcGly Heterointerface: Alignment of Energy Levels J. Phys. Chem. B 106 (2002) 5814–5819

[14] N. Masahashi, M.Oku, Superhydrophilicity and XPS study of boron-doped TiO2, Appl. Surf. Sci. 254 (2008) 7056–7060

[15] S. Akira, T. Masahiko, XPS and STM study of Nb-Doped TiO2(110)-(1 x 1) Surfaces, J. Phys. Chem. C. 117 (2013) 17680–17686.

[16] Amir Abidov, Bunyod Allabergenov, Jeonghwan Lee, Heung-Woo Jeon, Soon-Wook Jeong, and Sungjin Kim, X-Ray Photoelectron Spectroscopy Characterization of Fe Doped TiO2 Photocatalyst, IJMMM, 1, 3 (2013) 294–296.

[17] Olivier Rosseler, Mohamad Sleiman, V. Nahuel Montesinos, Andrey Shavorskiy, Valerie Keller, Nicolas Keller, Marta I. Litter, Hendrik Bluhm, Miquel Salmeron, and Hugo Destaillats, chemistry of NOx on TiO2 Surfaces Studied by Ambient Pressure XPS: Products, Effect of UV Irradiation, Water, and Coadsorbed K+, J. Phys. Chem. Lett. 4 (2013) 536–541.

[18] Ningdong Feng, Qiang Wang, Anmin Zheng, Zhengfeng Zhang, Jie Fan, Shang-Bin Liu, Jean-Paul Amoureux, and Feng Deng, Understanding the High Photocatalytic Activity of (B, Ag)-Codoped TiO2 under Solar-Light Irradiation with XPS, Solid-State NMR, and DFT Calculations, J. Am. Chem. Soc. 135 (2013) 1607–1616.

[19] Soo-Kyoung Kim, Min-Kyu Son, Songyi Park, Myeong-Soo Jeong, Kandasamy Prabakar, Hee-Je Kim, H. J. Kim, Surface modification on TiO2 nanoparticles in CdS/CdSe Quantum Dot-sensitized Solar Cell, Electrochimica. Acta. 118 (2014) 118–123.

[20] Xuemei, Zhou, Ning Liu, Patrik Schmuki, Ar+-ion bombardment of TiO2 nanotubes creates co-catalytic effect for photocatalytic open circuit hydrogen evolution, Electrochem. Commun. 49 (2014) 60–64.

[21] Wenjie Zhong, Shangbin Sang, Yingying Liu, Qiumei Wu, Kaiyu Liu, Hongtao Liu, Electrochemically conductive treatment of TiO2 nanotube arrays in AlCl3 aqueous solution for supercapacitors, J. Power Sources, 294 (2015) 216–222.